22.5 A 40GOPS 250mW Massively Parallel Processor Based on Matrix Architecture

Masami Nakajima[1], Hideyuki Noda[1], Katsumi Dosaka[1], Kiyoshi Nakata[1], Motoki Higashida[1], Osamu Yamamoto[1], Katsuya Mizumoto[1], Hiroyuki Kondo[1], Yukihiko Shimazu[1], Kazutami Arimoto[1], Kazunori Saitoh[1], Toru Shimizu[2]

[1]Renesas Technology, Itami, Japan
[2]Renesas Technology, Kodaira, Japan

Today's multimedia applications require CODEC processing, image processing, and image recognition with programmable flexibility. In particular, large-scale sum of absolute difference (SAD), convolution, and fast Fourier transform (FFT), are key elements for high-speed execution. On the other hand, advanced silicon processes today no longer permit power supply voltages to be scaled as before. Hence, energy efficiency will become the most important factor in SoC design. To meet the above demands, a massively parallel processor has been developed based on the Matrix architecture. It is called the MaTriX Processing Engine, or MTX for short. This architecture achieves 40GOPS of 16b additions at 200MHz clock frequency and 250mW power dissipation. 1Mb SRAM for data registers and 2,048 2b processing elements (PE) connected by a flexible switching network are integrated in 3.1mm^2 in 90nm low-power CMOS. The energy-efficient Matrix architecture supports 2,048-way parallel operations and the programmable functions required for multimedia SoCs.

Figure 22.5.1 shows a block diagram of the Matrix architecture. The components of this architecture are two data registers of 2,048×256b, and 2,048 2b PEs connected directly with data registers, and a controller with an instruction memory. The data register is composed of SRAM cells, Vertical Channel (V-ch) connecting PEs, and Horizontal Channel (H-ch) connecting memory cells and PEs. V-ch, made by the flexible switching network, generates the communication path between the PE and the power of 2 distance away PE in one cycle. V-ch can also be programmed at every cycle. As for H-ch, 3 operations (2 read, 1 write) are achieved by 2-bank composition with read-modify-write operation in one cycle. The Matrix architecture achieves the maximum performance at 200MHz V-ch/H-ch operations and 816Gb/s bandwidth with simple hardware.

Figure 22.5.2 shows the physical overview of the chip. It consists of 32 banks of hardware. A bank is composed of 64 entries, while each entry has double-sided 512b memory cells (256 bits on a single side), 4 sets of sense amplifiers (SA) / write drivers (WD), and PE. Adjacent banks are combined physically by sharing the SRAM memory cell array. Adjacent bit-line pairs belong to different banks, so the layout pitch of SA, WD, and PE is doubled. As a result, the inputs and the outputs of the memory cell array directly connect to PEs.

Figure 22.5.3 shows the detailed operation of V-ch. Figure 22.5.3(a) shows the operation flow of a 4-point FFT. In step 1, data is communicated between entries placed next to each other while simultaneously executing addition and subtraction with the data register. In step 2, data is communicated respectively between entries of a different distance. V-ch is required for operations between different entries. Using V-ch, 2b data of all entries can be transmitted in one cycle. Figure 22.5.3(b) shows a physical overview of the V-ch wire implementation. The MTX supports several types of shift steps. Though increasing the types of steps leads to an area overhead, symmetrical layout property and multi-layer-technology realize an area-efficient V-ch network. Figure 22.5.3(c) shows the V-ch circuit. Each PE broadcasts its X-register output via VCH_OUT signal by enabling VCH_EN signal. Each PE is also equipped with a V-ch input switch circuit which selects V-ch outputs from other PEs.

Figure 22.5.4 shows the detailed operation of H-ch. Figure 22.5.4(a) shows the overview of an entry operation with PEs and SRAM memory cells. PE in the MTX is composed of a 2b-ALU, a valid flag, and two X-registers, which temporarily store the data in memory cells. The ALU has a control circuit for high-speed multiplication by reducing partial products based on Booth's algorithm. Because the PE is based on a 2b ALU, no additional circuitry for high-speed operation (carry-look-ahead circuit for example) is required. The valid flag allows conditional execution. Therefore, a data dependent branch operation is supported even though the MTX is a SIMD processor. To enhance the area-efficiency, the SRAM memory cells employed in the MTX are single-ported. Figure 22.5.4(b) shows the operation flow diagram of an addition operation. As the ALU is 2b-grained, a data (an 8b data for example) stored in memory cells are processed in a bit-serial way. At cycle k, 2 bits (LSB) of DATA_B are read-out and they are stored in the X-registers at the next cycle k+1. At this cycle, 2 bits (LSB) of DATA_A are read-out from the other side of the memory array and added with the stored data of X-registers by the ALU. In addition, the output data of the ALU are written back to the memory cells in the same cycle. This design method is based on a read-modify-write operation of the SRAM with an asynchronous timing control. This design method can significantly reduce the size of a PE by eliminating unnecessary pipeline registers. Figure 22.5.4(c) shows the simulated waveforms of the proposed read-modify-write operation of the SRAM with an asynchronous timing control of activating word-line (WL), sense amplifier enable (SAE) and write driver enable (WE) signals one after another. With these results, when 2,048 sets of 16b additions are executed with 2,048 entries in parallel, the MTX can process all the data in 10 cycles. This is equivalent to an estimated processing performance of 40GOPS at 200MHz operation.

Figure 22.5.5(a) shows a chip micrograph of the MTX. 2,048 PEs and 1Mb internal SRAM are integrated in 3.1mm^2. A 200MHz operation at 1.2V is experimentally verified with a shmoo plot under the conditions of consecutive additions as shown in Fig. 22.5.5(b). Also, the maximum power consumption is experimentally evaluated as 250mW at 1.2V, 27°C. The features of the MTX is listed in Fig. 22.5.5(c). The application benchmark shows the performance of FIR filtering and FFT, which are key elements of multimedia applications, is adequately high, even though the MTX is a programmable processor.

Figure 22.5.6(a) shows energy efficiency expressed by two parameters: GOPS/mm^2 and GOPS/W. These metrics show 70 and 13 times better energy efficiency, respectively, compared to a conventional in-house DSP. Figures 22.5.6(b) and 22.5.6(c) show implementation area and power consumption in each part of the MTX. Figure 22.5.7 shows the energy efficiency compared with conventional parallel processors [1][2][3]. The processing capability per unit power consumption (GOPS/W) of the MTX is more than 20 times higher than other previous designs for 16b fixed-point additions.

References:

[1] B.Flachs et al., "A Streaming Processor Unit for a CELL Processor," *ISSCC Dig. Tech. Papers*, pp. 134-135, Feb. 2005.
[2] A.A.Bright et al., "Creating the BlueGene/L Supercomputer from Low-Power SoC ASICs.," *ISSCC Dig. Tech. Papers*, pp. 188-189, Feb., 2005.
[3] http://www.es.jamstec.go.jp/esc/eng/Hardware/arithmetic.html

 1-4244-0079-1/06/$20.00 ©2006 IEEE

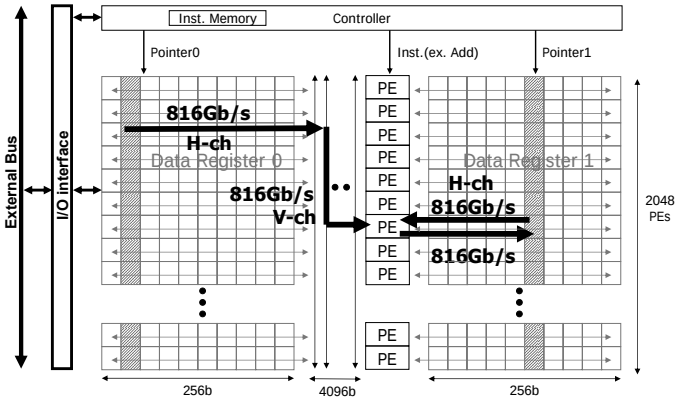


Figure 22.5.1: Block diagram of matrix architecture.

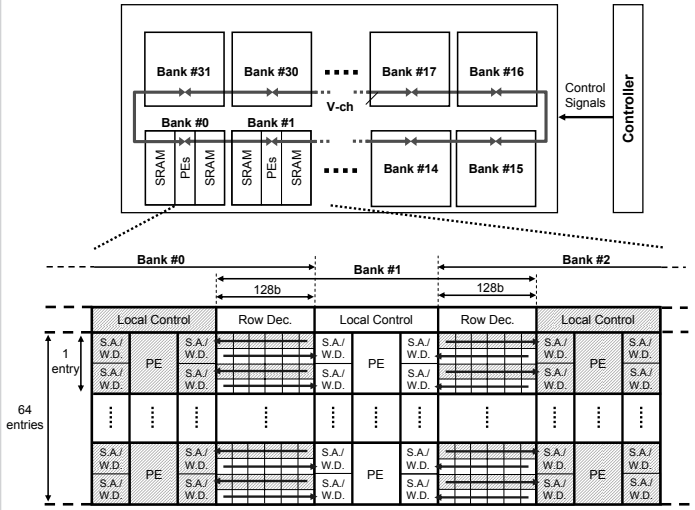


Figure 22.5.2: Physical overview of the chip.

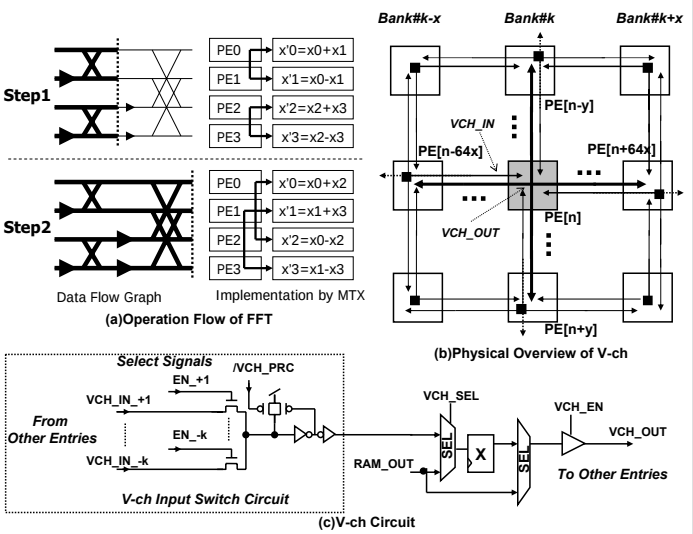


Figure 22.5.3: Operation of V-ch.

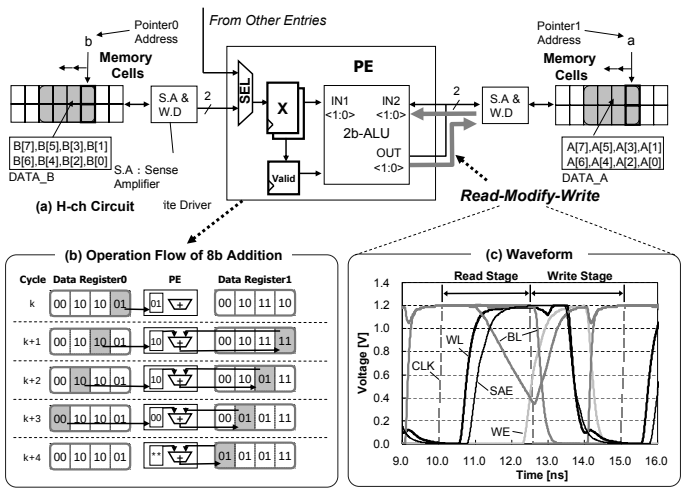
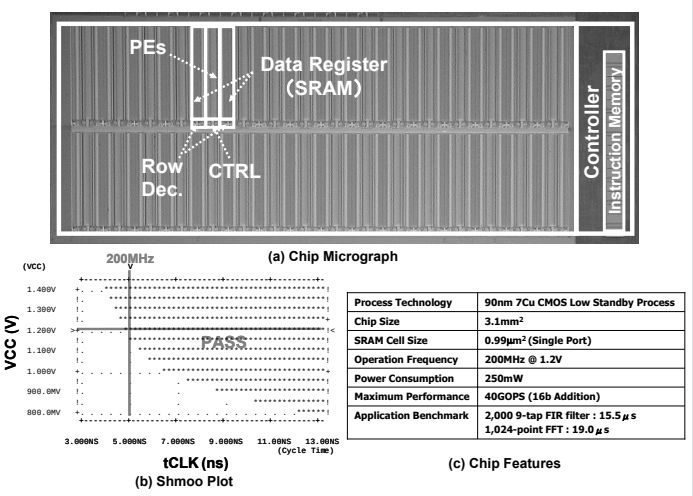


Figure 22.5.4: Operation of H-ch.



Process Technology	90nm 7Cu CMOS Low Standby Process
Chip Size	3.1mm²
SRAM Cell Size	0.99μm² (Single Port)
Operation Frequency	200MHz @ 1.2V
Maximum Performance	40GOPS (16b Addition)
Application Benchmark	2,000 9-tap FIR filter : 15.5μs 1,024-point FFT : 19.0μs

(c) Chip Features

Figure 22.5.5: Chip evaluation.

	GOPS/mm²	GOPS/W
MTX [This Work]	12.9	160
Conv. DSP [In-House]	0.183	13.3

(a) Comparison with Conv. DSP

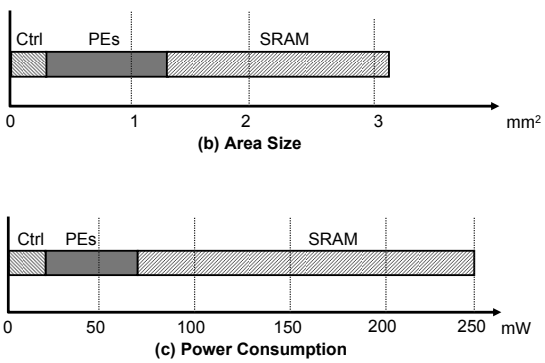


Figure 22.5.6: Comparison of energy efficiency.

Continued on Page 662

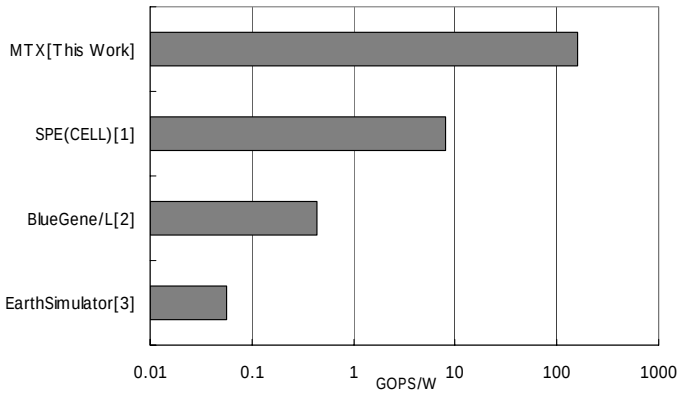


Figure 22.5.7: Comparison of processing capability per unit power consumption.

 1-4244-0079-1/06/$20.00 ©2006 IEEE